特　集「情報セキュリティとAI」

インターネットを対象としたデータマイニング技術の活用

Mining Internet Data

吉田　健一
Kenichi Yoshida
筑波大学大学院ビジネス科学研究科
Graduate School of Business Science, University of Tsukuba.
yoshida@gssm.otsuka.tsukuba.ac.jp, http://www2.gssm.otsuka.tsukuba.ac.jp/staff/yoshida/index.html

上原　宏
Hiroshi Uehara
（同　上）
ueharah@nttdocomo.co.jp

Keywords: data mining, internet, spam.

1. はじめに

インターネットを使った各種販売促進（販促）技術が新しい価値を創出しようとしている一方，迷惑メールやインターネットウイルスなどが新たなマイナスの社会要因をつくりつつある．本稿では，このような社会背景から望まれるデータマイニング技術の一つとしてキャッシュベースのマイニング技術について述べる．キャッシュベースのマイニング技術は，若干の精度の低下とひきかえに小容量のキャッシュメモリを使いながら高速に大量のデータから規則性を抽出しようとする技術である．本稿では上記の精度低下が実データの特徴を考慮すれば実用上許容範囲に収まることをSPAM filterなどへの応用を例に示す．

以下，2章でインターネットを使った各種販促技術の発展が創出しようとする新しい価値と，そのマイナス面について述べ，3章で迷惑メールなどマイナス面への現状の対策について述べ，4章でキャッシュベースのマイニング技術について説明する．

2. インターネットマーケティングとそのマイナス面

現在2011年の地上波デジタル放送への完全以降・インターネットを使った放送事業の胎動や，WWWを使った顧客管理などインターネットを使った各種販促技術の急速な発展に伴い，年間2兆円の市場規模をもつテレビ（以下TV）広告事業・放送業界の事業形態が変容しようとしている．例えば，今までは人手による視聴率調査以外に有効な方法のなかったTV広告の効果計測が，インターネット掲示板やblogの解析により，実時間で，かつ，個々の出演者に対する好感度調査なども同時に実施する形で行える可能性が生じつつある．このことは商業活動への影響を通して社会に大きな影響を与えることが予想される．

この背景には，インターネットを利用する人々が国民の約6割強に達し，また消費行動の中で重要な位置を占めつつあること[白書 05]がある．また，インターネット上の各種データを解析する情報処理技術の進展も重要な意味をもっている．本章では，インターネットを使ったマーケティングの一例とマイナス面のもたらす課題について述べる．

2・1　インターネット掲示板を使ったマーケティング

インターネットの浸透がマーケティングに及ぼす好例として，2ちゃんねると呼ばれるインターネット掲示板と，そこに投稿される記事の内容を分析することによるTV視聴率の実時間解析技術がある．

上原[Uehara 06]は実況スレッドと呼ばれる2ちゃんねるの記事が，放送中の人気ドラマ番組に対して，1時間当たり数千という膨大な数になることに着目し，キーワードの単純な出現回数計測だけで，特定の俳優への分単位での視聴者の興味の変化が分析できることを報告している．例えば図1は，掲示板から抽出した出演者に対

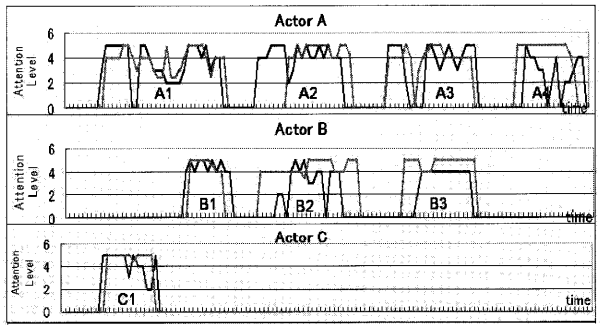

図1　2ちゃんねるの解析によるTVドラマ視聴の分析[Uehara 06]

する1分ごとの興味の変化（黒い線）と，アンケート調査で調べた興味の変化（灰色の線）が一致している様子を示している．図1から明らかなように，掲示板から簡単に取れる情報は，今まで入手するにはアンケートの実施など人手が必要であった情報と精度良く一致する．TV広告が産業に及ぼす影響を考えた場合，このことの応用価値・社会的意義は大きい．

このような解析を可能にした背景には，若者を中心に，2ちゃんねるに代表されるインターネット掲示板を見ながらTVを視聴し，視聴と同時にTV番組に対するコメントを掲示板に投稿するTV視聴者が増加していることがある．この結果として，インターネット掲示板の情報量が増えており，従来のような複雑な自然言語解析技術を用いなくても単純な統計的な処理だけで，番組出演者に対する興味を抽出することができるようになってきている．このことは

- 地上波アナログ放送の視聴率のように，従来は人手を要していた情報の入手が，人手をかけずに実時間で計測できる．
- 上記は，単純な視聴率調査にとどまらず，分単位でどの出演者に注目が集まっているかという，時間単位においても，解析単位（この場合は出演者）においても詳細な情報の解析となっている．
- 2ちゃんねるは日本特有のインターネット掲示板であるが，blogなど類似のWeb技術を用いたTVとインターネット掲示板の同時視聴の形態は，若い人を中心に世界的に広がる可能性がある．この場合，関連する広告業界などへの社会的影響は世界的に見ても大きい．

といった観点から関連研究分野への大きな影響が予想できる．

また，同様な研究は，スポーツ番組を対象とした宮森の研究[宮森 04]や，ニュース番組を対象とした上原の研究（図2は[上原 05]に掲載されたニュース番組の解析事例）など事例を増やしつつあり，インターネットの社会への浸透を背景に，新しいマーケティング技術として重要性を増しつつある．

これらは，単純なデータマイニング技術であっても，

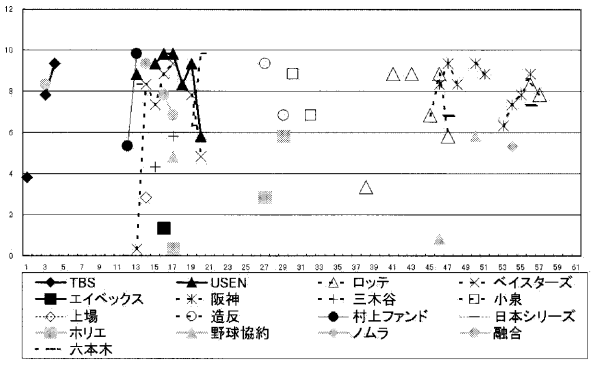

図2　2ちゃんねるの解析によるニュース視聴の分析[上原 05]

インターネットのもつ強力な情報収集能力と組み合わせることにより，インターネット時代の商業広告を支える情報処理技術，マーケティング技術になり得ることを示唆している．

2·2　インターネットセキュリティの課題

上記のようなインターネット時代の商業広告をささえる情報処理技術やマーケティング技術の研究が進む一方，広告に関連して，行き過ぎた技術利用による迷惑メールが社会問題化している[総務]．道徳的に許されるか否かの判断は別にして，迷惑メールを広告技術として見た場合

- 名簿作成専業業者が存在するなど，潜在顧客リストの自動作成の仕組みが確立している．
- 安価に宣伝文を配布可能である．
- 反応してきた有望顧客リストの自動作成と名簿管理も，自動化されている．
- 広告に対する反応が速い．

と非常に優れた性質をもっている．このことは，もともと研究者間の実験ネットワークとして設計された経緯をもつ，性善説に基づくインフラ（TCP/IP）技術と，やはり研究者間の連絡のために設計されたアプリケーションであるWebやメールの特質とも重なり，迷惑メールの対策が進まないことの一因にもなっている．

またあくなき利便性の追及はシステムの複雑化を招き，セキュリティホールを増やす結果につながっている．例えば商用・非商用を問わず，使い勝手の良いWebブラウザにセキュリティホールが見つかるという事例は良く発生する．新種のインターネットウイルスも日常的に報告・警告されている．

2·3　データマイニング技術への期待

前述のような迷惑メールやインターネットウイルスの中にはデータマイニング技術を応用することにより，比較的容易に対策できるものもある．例えば，迷惑メールに関して「類似したメールの数を数え一定数以上のものをスパムと分類する」という単純な頻出アイテム検出のアイディアが有効なことが実証的に示されている[*1]．同様なアイディアはインターネットセキュリティの分野では各所に使用できると思われる（図3）．

例えばインターネットウイルスは頻出するsource IP addressとdestination port numberの組合せをバックボーンに流れるパケットの情報から抽出することで検出可能に思われるし，分散DoS攻撃の検出は特定の計算機に多数の計算機から多量のsynパケットが送られていることを検知できれば検出可能に思われる．セキュリティ向上の観点から，P2Pによるネットワークの使用状況を把握しておくことはネットワークの重要な管理業務の

*1 4·1節で上記研究[Yoshida 04]について概説する．

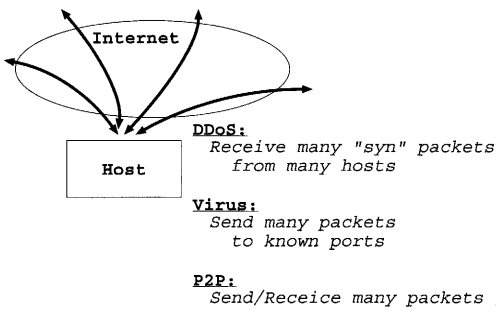

図3 ウイルス，分散DoS，P2Pの検出

一部であるが，これも特定のsource IP adressとsource port numberの組合せとして検出可能に思われる．

これらはいずれも単純な頻出アイテムまたはアイテムの組合せの検出技術が迷惑メールやインターネットウイルスなどのマイナス面など新たなマイナスの社会要因への対応技術として有望なことを示唆し，データマイニングへの期待となっている．

2・4 データマイニングの研究課題

前述のように，迷惑メールやインターネットウイルスなどの新たなマイナスの社会要因を防ぐ手段として，データマイニングの技術は有望と考えるが，いくつかの技術的なハードルが残されている．例えば，迷惑メールやインターネットウイルスなどの対策を考えたときには，大量のオンラインデータからリアルタイムでデータマイニング [Agrawal 94] を行おうとする研究 [Hidber 99] が重要である．特に最近の研究で，Webやスパムメールなど Zipf の法則に従う現象が多いことがわかってきた（例えば [Nishikawa 98, Yoshida 04]）が，そのような性質をもつ大量のデータからの規則性抽出は重要な研究テーマである．

例えば，インターネットバックボーンに流れるパケットデータから頻出パターンをマイニング技術を使って発見することはインターネットウイルスや分散DoS攻撃を検出することになり，社会リスク低減を目的とした技術としての価値が高いが，このようなデータは非常に大量であり，高速に生成されるため，単純に従来のマイニング技術を使っただけでは解析できない．表1に1 Gbpsのアクセス回線に流れるネットワークパケットのデータ量を示す．実際にパケットデータを代表的な頻出アイテムの組合せ抽出プログラム LCM-v2[Uno 04] で解析を試みた場合，40 Mパケットの処理に約1.8 GByteのメモリ容量を必要とし，かつ，パケット数と必要なメモリ容量が比例関係にあった．これは従来技術で解析可能なパケットはおおよそ40秒分が最大であることを意味する．

ネットワークの保守管理にはこの種の解析を1時間単位で実施することが求められることが普通であり，現状の頻出アイテムの組合せ抽出プログラムでは事実上解析することはできない．このことはネットワーク研究者の興味の対象にもなっており，解析対象をネットワークデータに限定しても各種の研究（例えば，[Duffield 03, Estan 02, Golab 03, Hohn 03, 石橋 06, Kumar 04, Mori 04, 森 06]）が行われているが，いまだ研究の余地が残っている．

表1 ネットワークパケットのデータ量

観測期間	パケット数	パケット種類
1秒	1M	1K
1分	60M	4M
1時間	4G	210M
1日	86G	4G
1週	605G	35G

3. 現状のインターネットセキュリティ技術

本章では初めに現在主流のインターネットのセキュリティに関する二つの技術を解説し，それぞれの問題点を指摘する．

3・1 SPAM filter

現在主流の迷惑メールの対策システム（いわゆるスパムフィルタ）は，メールに含まれる単語の出現頻度情報をもとに，迷惑メールに固有の単語の出現頻度パターンを機械学習の手法などを使い学習し，識別している（例えば [Graham 02, Graham 03, Robinson 02] など）．すなわち，具体的な商品名や「完全」「無料」といった単語を多く含むものは迷惑メールであると推定するが，そうでないものは迷惑メールではないと推定するという考えである．

最も知られた方法の一つである [Graham 02] は，

- はじめに迷惑メールとそうでない普通のメールを収集し，
- 次に各単語ごとに，その単語を含むメールが迷惑メールであるか否かの確率を，上記収集メール群から計算しデータベース化しておき，
- 新しいメールがきたときに，メールが含む単語と上記で計算した単語ごとの確率を用いて，メールが迷惑メールか否かの確率を計算している．

[Graham 02] では，上記で第2，第3ステップにおいて通常の統計的な計算方法を使わず経験的に修正した方法のほうが識別性能が上げられることが主張されているが，基本的には統計的な考えを参考にしている．

しかしながら，迷惑メールを配布しようとする者達は，このようなフィルタ技術の裏をかくために，不要な単語を挿入するなど日々工夫を行っている．このことが単語の出現頻度情報だけをもとに迷惑メールを普通のメールと識別する問題の本質的な難しさとあいまって，識別精度の向上を阻んでいる．すなわち「無料」といった単語は迷惑メール以外の普通のメールにも使用される単語であり，撹乱のために挿入された不要な単語とともに誤判定率が上がる原因となる．また

- 新しい商品名（単語）などについて，その単語を含

むメールが迷惑メールであるか否かの確率を，日々更新しないと識別性能が急速に下がっていく．このため，利用者が頻繁に確率を記憶したデータベースの更新処理を行う必要がある．

- 迷惑メールは重複が多く，機械学習システムの識別性能を計るために一般的に用いられている 10 fold validation は，テストセットとトレーニングセットが同じデータセットになりやすい．この事実を無視した事前の性能評価は実運用時に予定より低い性能となって現れ，不信を招いた結果，その後の利用停止を招きやすい．

などの事情とともにスパムフィルタが迷惑メール対策の切札にならない理由の一つとなっている．

3·2 ネットワークトラフィック分析

前述のように，インターネットバックボーンに流れるパケットデータから頻出パターンをマイニング技術を使って発見することはインターネットウイルスや分散 DoS 攻撃を検出することに利用できる可能性があり，種々研究が進んでいる．

図 4 は最も基本的な Sliding Window 法 [Golab 03] によるパケット数の多いフローの解析アイディアを示したものである．

Sliding Window 法ではインターネットバックボーンに流れるパケットデータは，最新の Basic Window N の情報のみが計算機のメモリに記憶される．古いパケットの情報は過去の Basic Window ごとに頻出アイテムの情報のみメモリに保持され，パケットデータ自体は廃棄される．Window ごとの頻出アイテムの情報も LRU アルゴリズムに従い古いものから順に廃棄され，メモリ上に保持されている Window ごとの頻出アイテムの情報から全体の頻出アイテムの情報，すなわちパケットの多いフローの情報が計算される．

通常，頻出アイテムを記憶するために必要なメモリ容量はパケットデータの情報をそのまま記憶する Basic Window が必要とするメモリ容量より大幅に小さいため，Sliding Window 法に必要なメモリ容量はおおよそ Basic Window 1 個分である．この性質を利用し，適切な Basic Window のサイズと，ある程度大きな Basic Window の数を選ぶことによる，フローの解析が試みられている．しかし，パケットデータに含まれるフローの分布はいわゆる Zipf の法則に則っており，Basic Window は少数しかパケットを含まない非常に多数のフローの情報で占められることになる．このため，さらにメモリ効率の良い手法の開発が必要である．

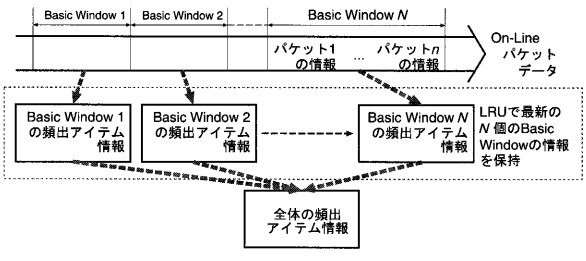

図 4 Sliding Window 法によるパケット数の多いフローの解析

また，インターネットウイルスの発見には単純なアイテムの検出ではなく，アイテムの頻出する組合せを解析する必要があるが，オンラインで多量のデータから頻出アイテムの組合せを効率良く発見する手法についても研究の余地が残っている．

4. キャッシュベースのマイニング技術

本章では，前章での議論を背景に，著者らが研究しているキャッシュベースのマイニング技術について解説する．

4·1 SPAM filter での利用

キャッシュベースのマイニング技術は迷惑メールの対策システム [Yoshida 04] を開発した過程で考案したものである．図 5 に著者らが開発した迷惑メールの対策システムの基本アイディアを示す．

著者らが開発した迷惑メールの対策システムは，従来技術とは異なり，単語の出現情報は「類似メールが何通あるか」の判断だけに用いて，スパムか否かの判断は「同じメールが多数あればスパム」（図 5 において出現頻度の高い部分がスパム）という単純な基準を用いている．アイディアは単純であるが，迷惑メールを配布しようとする者達には「宣伝のための同じ情報を多量に配布しなければ商売にならない」という制約があるため，従来技術（表 3）よりも識別精度のうえでも，処理速度のうえ

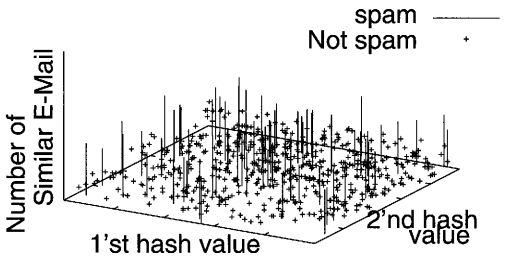

図 5 類似メールの出現頻度

表 2 実験結果（[Yoshida 04] より抜粋）

e-mail の総数	53,985,002
spam と判定されたe-mail の数	12,324,762
メモリ容量	825 MByte
CPU 時間	4340 sec
再現率	100 %
精度	100 %

表 3 従来技術との比較

	SVM	NB	C4.5	Knn	[BF 03]	[SA 03]
再現率(%)	81	47	77	81	73	83
精度(%)	99	97	95	100	98	22
速度比(倍)	1009	106	379	14528	87	300

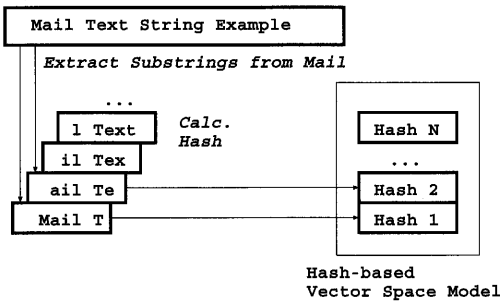

図6 部分文字列のハッシュ値を使った文章表現

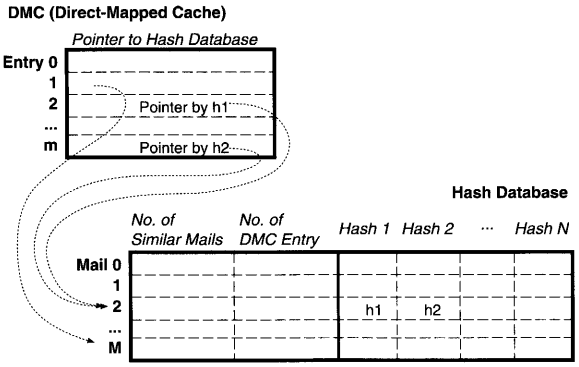

図7 類似メール記憶のためのキャッシュ構造

でも優れた結果（表2）を得ている．すなわち表に示したように，単語の出現頻度のみで識別処理をしようとする従来の迷惑メール対策システムで再現率・精度ともに100%のものはなく，キャッシュベースの手法は速度的にも87倍以上高速である．

上記対策システムは「同じメールが多数あればスパム」という単純な基準により文章の表現自体もTF・IDFのような，ある程度複雑な処理を使ったものでなく，複数の部分文字列を切り出し各部分文字列のハッシュ値を使うもの（図6）で十分な性能を達成できた．一方，1日当たり数億通のメールの中から類似メールを効率良く検索するためには，全メールのデータ量に比較すればはるかに小さなメモリの中に，頻出する文字列パターンを効率良く記憶し，高速に検索する仕組みの開発が必要であった．

図7に上記対策システムが類似メールを記憶・検索するために利用しているキャッシュ構造を示す．図7下段のHash Databaseはメールごとに各部分文字列のハッシュ値を記憶しておくためのキャッシュ構造である．いかに類似したものの多いメールの情報をHash Databaseの中に記憶しておくかの制御が，対策システムの性能を決めている．

図7上段のDirect-Mapped Cacheが，この制御と高速検索のために設けられたキャッシュ構造である．Hash Databaseに新しいメールの情報を記憶するときに，そのハッシュ値の一部を使ってDirect-Mapped Cacheから，新しいメールのHash Database上の記憶位置にポインタを作成しておく．到着したメールに似たものがないかの検索は，到着したメールから作成したハッシュ値でDirect-Mapped Cache経由でHash Databaseを調べることで行う．類似したメールはハッシュ値に同じものを含むので，すでに受け取ったメールに類似したものがあれば，この仕組みにより，高速に検索可能である．

Direct-Mapped Cacheから，Hash Database上の記憶位置にポインタを作成するときには古いポインタを上書きして廃棄する．したがってDirect-Mapped Cache上のポインタは一見ランダムに削除されるが，到着したメールとHash Databaseに記憶されているメールが類似していると判定されたときに，Hash Databaseに記憶されているメールに対するDirect-Mapped Cacheからのポインタを再作成することで，頻出するメールの情報がうまく保存されることが実験により確認されている[Yoshida 04].

4·2 一般的マイニング問題への適用

上記の迷惑メール対策システムの経験から，著者らは大量のオンラインデータから小容量のメモリを使ってリアルタイムで頻出するアイテムの組合せを抽出する方法として，固定サイズのキャッシュを使う方法を検討している[吉田 05]．「頻出するアイテムの組合せ」ではなく，「頻出するアイテム」の抽出に限定すればアイディアは単純であり，固定サイズのキャッシュを使って処理対象のデータに含まれるアイテムの数を数え，設定値以上の回数出現したアイテムを頻出アイテムとして出力する（図8に疑似コードを示す）．

頻出するアイテムの組合せを検出する場合，キャッシュの構造が若干複雑になるが，基本的には同じ考え方が使える．すなわち，初期は頻出アイテムの抽出を行っておき，頻出すると判断されるアイテムが見つかった後はその頻出するアイテムを含む二つのアイテムの組合せについても出現回数をチェックする．N個のアイテムの組合せが頻出すると判断された後は，そのN個のアイテムの組合せを含む$N+1$個のアイテムの出現回数をチェックする．正確には処理速度を上げるための工夫などが必要となるが，ここでは詳細は割愛する．

キャッシュの容量が十分大きい場合，アイテムの記憶位置を決める処理（図8の3行目）は単純である（過去に出現して記憶されているものはハッシュ関数などを用いて記憶位置を決定し，新しいアイテムであれば空きエリアを記憶位置とする）が，容量が全アイテムを記憶するだけ確保できない場合，何らかの指標に基づきすでに

```
Create empty heap;
while (input item) do
  i = index of item in heap;
  increment heap_cnt[i] by 1;
  if (heap_cnt[i]>thresh_hold)
    print message;
done
```

図8 キャッシュベースのマイニング技術：基本アルゴリズム

```
int i = random() % HEAP;
for (p=1; (heap_cnt[i]>p); p++)
   i = random() % HEAP;
```

図9 Random2: C プログラムコード

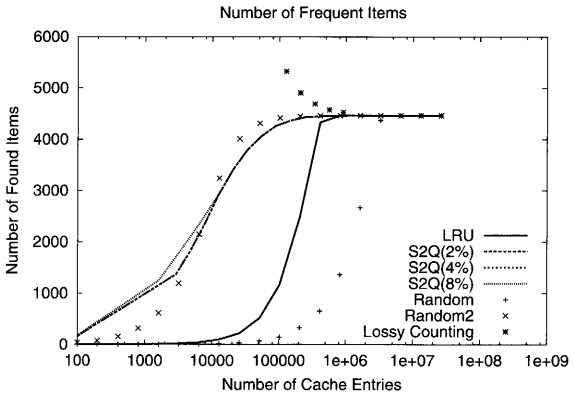

図10 キャッシュベースのマイニング技術:性能評価の結果 [吉田 05]

あるアイテムのデータを削除して，キャッシュメモリを再利用することを考える．

未使用期間が最も長いエントリを削除する LRU (Least Recently Used) アルゴリズムは，このようなときに用いられる標準的な方法であるが，実験では LRU に比べて単純な Random2（図9）の性能が良いことがわかってきた [吉田 05]．Random2 は前述の迷惑メール対策システムのキャッシュ管理方式と同じく，古いエントリを上書きして新しいデータを記憶し，データを一見ランダムに削除するが，複数回出現したデータをなるべく捨てないような処理をしている．図9に C による Random2 のプログラムコードを示す．1行目は Random そのもの，2行目と3行目は複数回出現したデータをなるべく捨てないようにしている処理を実現しているが，基本的にはこの修正をした後も，乱数で決めた数をキャッシュサイズで割って余りが新しいアイテムの記憶位置（削除するデータの記憶位置）となっている．

図10にキャッシュベースのマイニング技術について性能を評価した結果を示す．[吉田 05] では代表的な既存手法である Lossy Counting[Manku 02] と，キャッシュベースのマイニング技術において，キャッシュの内容管理を LRU, Simplified 2Q, Random, Random2 で行ったものの性能を評価し，キャッシュの内容管理を Random2 か Simplified 2Q[Johnson 94] で行った場合のキャッシュベースのマイニング技術が性能の良いことを報告している．すなわち，図10で横軸はキャッシュ容量（エントリ数），縦軸は各手法が見つけた頻出アイテムの数を示しているが，Lossy Counting の性能が 10^5 を境に急速に落ちている*2のに比較して，キャッシュの内容管理を Random2 か Simplified 2Q で行った場合のキャッシュベースのマイニング技術は1桁近く小さなキャッシュ容量でも動作している．

以上，キャッシュベースのマイニング技術について説明してきた．キャッシュベースのマイニング技術のように比較的小容量のメモリを使い，大容量のデータの特徴を抽出する技術は，今後重要性を増していくと考える．

5. ま と め

新しい情報処理技術が新しい価値を創出しようとしている一方，迷惑メールやインターネットウイルスなどのマイナス面が新なマイナスの社会要因をつくりつつある．本稿では，このような社会背景から望まれるデータマイニングの諸側面について討議を試み，

- 新しい情報処理技術が社会に及ぼしつつあるプラスの影響と
- 迷惑メールやインターネットウイルスなどのマイナス面とデータマイニング技術への期待
- 上記マイナス面に対する現在主流となっている対策技術の現状とその課題
- 比較的小容量のメモリを使い，大容量のデータの特徴を抽出する技術としてのキャッシュベースのマイニング技術

について述べた．

本稿では主に技術的側面について取り上げたが，新しい情報処理技術がもつマイナス面を対策するには，技術的な検討だけでは不十分である．例えば通信内容のマイニング処理は盗聴につながる側面をもっており，使用に際してはリスク対策が別の社会的に見たときに負の要因を生成しないか議論したうえで使っていく必要がある．どこまでの解析を社会的に容認するかについては一つの技術の正の側面・負の側面，両面から見たうえで社会の中でその功罪について議論し合意を形成していく必要がある．現状，技術的な開発に社会的な合意形成が追いついていっていないことも大きな問題となっていると考える．

*2 Lossy Counting は小さなキャッシュでは頻度を過剰評価するので見かけ上の頻出アイテムの数が誤認により増える．一方キャッシュベースのマイニング技術は見逃しにより頻出アイテムの数が減る．

◇ 参 考 文 献 ◇

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast algorithms for mining association rules, in Bocca, J. B., Jarke, M. and Zaniolo, C. eds., *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487-499, Morgan Kaufmann (1994)

[BF 03] http://www.h2.dion.ne.jp/nabeken/bsfilter/ (2003)

[Duffield 03] Duffield, N., Lund, C. and Thorup, M.: Estimating flow distributions from sampled flow statistics, *Proc. ACM SIGCOMM*, pp. 325-336, Karlsruhe, Germany (2003)

[Estan 02] Estan, C. and Varghese, G.: New directions in traffic measurement and accounting, *Proc. ACM SIGCOMM*, pp. 323-336, Pittsburg (2002)

[Golab 03] Golab, L., DeHaan, D., Demaine, E., Lopez-Ortiz, A., and Munro, J. I.: Identifying frequent items in sliding windows over on-line packet streams, *Proc. ACM SIGCOMM Internet*

Measurement Conference, Miami, USA（2003）
[Graham 02] Graham, P.: http://www.paulgraham.com/spam.html（2002）
[Graham 03] Graham, P.: Better Bayesian filtering, *Proc. 2003 Spam Conference*（2003）
[白書 05] 総務省平成17年版情報通信白書（2005）
[Hidber 99] Hidber, C.: Online association rule mining（1999）
[Hohn 03] Hohn, N. and Veitch, D.: Inverting sampled traffic, *Proc. ACM SIGCOMM Internet Measurement Conference, Miami*, USA（2003）
[石橋 06] 石橋圭介, 森 達哉, 川原亮一, 廣川 裕, 小林淳史, 山本公洋, 坂本仁明：異なり数上位Nホスト推定および異常検出への応用, 電子情報通信学会2006年総合大会（2006）
[Johnson 94] Johnson, T. and Shasha, D.: 2Q: a low overhead high performance buffer management replacement algorithm, *Proc. 12th Int. Conf. on Very Large Databases*, pp. 439-450, Santiago, Chile（1994）
[Kumar 04] Kumar, A., Xu, J., Wang, J., Spatscheck, O. and Li, L.: Space-code bloom filter for efficient per-flow traffic measurement, *Proc. IEEE Infocom*, Hong Kong（2004）
[Manku 02] Manku, G. and Motwani, R.: Approximate frequency counts over data streams, *Proc. 28th Int. Conf. Very Large Data Bases*, pp. 346-357, Hong Kong, China（2002）
[宮森 04] 宮森 恒, 中村聡史, 田中克己：番組実況チャットを利用した放送コンテンツの自動インデキシング, 信学会パターン認識・メディア理解研究会予稿, 第NLC2004-123巻（2004）
[Mori 04] Mori, T., Uchida, M., Kawahara, R., Pan, J., and Goto, S.: Identifying elephant ows through periodically sampled packets, *Proc. ACM SIGCOMM Internet Measurement Conference*, Taormina, Sicily, Italy（2004）
[森 06] 森 達哉, 石橋圭介, 上山憲昭, 川原亮一：NetHost: ホスト毎トラヒックサマリ集約方法の提案, 信学会2006年総合大会（2006）
[Nishikawa 98] Nishikawa, N., Mori, Y., Hosokawa, T., Yoshida, K. and Tsuji, H.: Memory-based architecture for distributed WWW caching proxy, *Proc. of World Wide Web Conference 98*, pp. 205-214（1998）
[Robinson 02] Robinson, G.: Spam Detection http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html（2002）
[SA 03] SpamAssassin http://useast.spamassassin.org/（2003）
[総務] http://www.soumu.go.jp/joho_tsusin/d_syohi/m_mail.html
[上原 05] 上原 宏：携帯端末とマスコミュニティ連携の可能性（平成17年度情報化月間行事ITシンポジウム）（2005）
[Uehara 06] Uehara, H. and Yoshida, K.: Automating viewers' side annotations on TV drama from internet bulletin boards, 情報処理学会論文誌, Vol. 47（2006）
[Uno 04] Uno, T., Kiyomi, M., and Arimura, H.: LCM ver.2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets, *IEEE ICDM' 04* Workshop FIMI' 04（2004）
[Yoshida 04] Yoshida, K., Adachi, F., Washio, T., Motoda, H., Homma, T., Nakashima, A., Fujikawa, H., and Yamazaki, K.: Density-based spam detector, 電子情報通信学会英文誌（D）, Vol. E87-D（2004）
[吉田 05] 吉田健一, 勝野 聡, 藤田昌克, 鶴 正人, 阿野茂浩, 山崎克之：キャッシュを使った頻出アイテムの抽出, 信学誌, Vol. J88-B, No. 10（2005）

2006年6月15日受理

著 者 紹 介

吉田 健一（正会員）

1980年東京工業大学理学部情報科学科卒業，同年，（株）日立製作所入社．1992年9月博士（工学，大阪大学）．2002年より筑波大学大学院ビジネス科学研究科教授．インターネット上の各種データを，機械学習の手法を使って解析する研究に従事．情報処理学会会員．

上原 宏

1984年横浜国立大学経済学部卒業，1997年青山学院大学国政政治経済学研究科修了．2006年筑波大学ビジネス科学研究科修了．現在，（株）NTTドコモマルチメディアサービス部勤務．インターネットコミュニティを利用したテキストマイニングに興味をもつ．情報処理学会会員．